REMARKS

Applicants gratefully acknowledge Examiner Do for taking time from his busy schedule to conduct a telephone interview on March 14, 2007, with co-inventor Dr. Gustavson and Applicants' representative F. Cooperrider. It is believed that this interview was very beneficial because it allowed the co-inventor to explain the significance of the present invention and its distinctions from the cited prior art. Examiner Do seemed to agree that there were indeed differences but was unwilling to agree that the original claims adequately articulated these differences.

Claims 1-20 are all the claims presently pending in the application. Various claims have been amended to more particularly define the invention.

It is noted that the claim amendments are made only for more particularly pointing out the invention, and not for distinguishing the invention over the prior art. Further, Applicant specifically states that no amendment to any claim herein should be construed as a disclaimer of any interest in or right to an equivalent of any element or feature of the amended claim.

Claims 1-20 stand rejected under 35 U.S.C. § 101 as allegedly directed to non-statutory subject matter. Claims 1-5, 9, 14, and 15 stand rejected under 35 U.S.C. § 102(a) as anticipated by U.S. Patent No. 6,601,080 to Garg. Claims 7, 8, 11, 12, 16, 17, 19, and 20 stand rejected under 35 U.S.C. § 103(a) as unpatentable over Garg, further in view of non-patent literature by Philip Alpatov, et al.

These rejections are respectfully traversed in the following discussion.

I. THE CLAIMED INVENTION

The claimed invention, as exemplarily defined in independent claim 1, is directed to a method of executing a linear algebra subroutine on a computer having at least one cache. Data from two of three matrices involved in processing a linear algebra subroutine is streamed to the first matrix, such that submatrix data of the two matrices residing in higher level cache or memory is streamed to submatrix data of the first matrix residing in the cache. The streaming provides data from the higher level to the data in the cache as required for the correct processing for the executing of a linear algebra subroutine.

The present inventors have recognized that conventional linear algebra processing

based on LAPACK subroutines, for example, are not optimal.

The claimed invention, on the other hand, along with various other techniques described in the co-pending applications, provides techniques that improve processing efficiency. More specifically, the present invention provides a memory management method allowing for a streaming of data through a cache, using another operand as having the "matrix role" and being resident in the cache.

II. THE 35 USC §101 REJECTION

Claims 1-20 stand rejected under 35 U.S.C. §101. As best understood, Examiner Do considers that all claims "... merely disclose steps of streaming data from cache without [regard] to any particular practical application or tangible result." Examiner Do also considers that claims 14-18 are directed to "signal medium" and, therefore, non statutory.

Applicants respectfully disagree.

First relative to claims 14-18, Applicants submit that these claims are clearly addressed to "[a] signal bearing medium tangibly embodying a program of machine-readable instructions executable by a digital processing apparatus ....." Such claims are often referred to as "Beauregard claims", after *In re Beauregard*, 53 F.3d 1583 (Fed. Cir.,1995), wherein the USPTO Commissioner conceded that such claims are indeed statutory subject matter. The Examiner might want to read this case holding, since it is only three paragraphs in length and might also want to check out the resultant patent, US Patent No. 5,710,578 to Beauregard, et al., issued on January 20, 1998.

Relative to claims 9-13, these claims are directed to an apparatus and are, therefore, clearly directed to a machine, one of the four enumerated categories specifically recited in 35 USC §101 as statutory subject matter.

Relative to the method claims, as discussed during the telephone interview, Applicants submit that the present invention does indeed have the prerequisite practical application and tangible result (wherein "tangible" means "real-world") because its method improves DGEMM performance in two ways: faster DGEMM kernel processing via streaming; and, providing only 4 ways to block for the entire DGEMM operation via the use of faster DGEMM kernels. Since these two benefits provide improved efficiency and speed in DGEMM processing on a computer, there clearly is a beneficial real world result.

Thus, from another perspective, the present invention can be viewed as a memory

management technique in a computer that improves efficiency for processing dense linear algebra subroutines.

The independent claims have been amended in an attempt to find wording more acceptable to Examiner Do.

In view of the foregoing, the Examiner is respectfully requested to reconsider and withdraw this rejection.

III. THE PRIOR ART REJECTIONS

The Examiner alleges that Garg teaches the claimed invention defined by claims 1-5, 9, 14, and 15 and, when modified by Alpatov, et al., renders obvious claims 7, 8, 11, 12, 16, 17, 19, and 20.

Applicants respectfully disagree and submit that there are elements of the claimed invention which are neither taught nor suggested by Garg.

The present invention provides a method and structure for producing high performance linear algebra routines using streaming. An exemplary feature of this invention is that it allows an L1 cache resident matrix to get essentially infinite reuse. Hence, the term "streaming" is used, although this choice of terminology may be unfortunate, since "streaming" has other connotations in this field. Also, by using streaming in accordance with the present invention, one can reduce the number of efficient ways to block data for matrix multiplication.

In contrast, Garg teaches a method for efficiently solving a system of equations involving a sparse positive definite symmetric matrix. Garg uses a supernodal approach to perform a CMOD operation on either a 1-D trapezoidal sparse representation or a 2-D representation of the supernode. When using a 2-D representation, standard prior-art cache blocking techniques are used. In this latter case, the CMOD operation is, in fact, used by Garg as a standard DGEMM operation. (Column 14, lines 33 to 36 of Garg)

The domain of the present invention is dense linear algebra (DLA) algorithms, whereas the domain of Garg is sparse linear algebra for only Cholesky factorization. Therefore, Applicants submit that the environment and purpose of Garg is clearly different from that of the present invention.

Relative to claims 1 and 2, Garg is using standard prior-art cache blocking techniques when it goes to 2-D processing for sparse Cholesky factorization via a supernodal approach.

In contrast, the present invention is concerned with a generalization of level 1 cache blocking which is applicable to almost all DLA algorithms. So, as mentioned in the above-mentioned telephone interview, Garg could use the present invention to improve his invention, but the converse is not true.

Thus, as described by the independent claims, the present invention establishes that one of the three matrices will serve in a "matrix role" for the subroutine and determines that submatrix data of this first matrix will be cache resident. The data of the remaining two matrices will then be streamed through cache from higher levels of memory and thus improve the efficiency of the subroutine.

In the rejection currently of record, the Examiner points to lines 38-55 of column 7 of Garg. However, Applicants submit that, whatever other similarities might be present in Garg relative to the method of the present invention, these lines clearly describe using L2 cache for the supernode storage and make no mention of streaming matrix data from other matrices from higher levels of memory. In fact, Garg does not discuss the factorization of the triangle part of the super-nodes and in particular the root super-node. Please refer to columns 11 and 12 starting at line 55. Also, to Figures 7 and 8. On lines 2 to 6 Garg is clearly using standard 2-D representations. Lines 15 to 26 of col 2 and Figure 8 show how Garg uses L2 standard L2 cache blocking with 3 source super-nodes of sizes 3, 4 and 6. They eventually become the large target super-node of size 13. These target super-nodes must be factored. Now consider the root super-node and refer to lines 56 to 59. This special node is therefore stored in full standard format even though Garg may use L2 blocking to update it. Eventually however, after all CMOD updating it must be Cholesky factored. This full triangle is stored in full format and hence wastes half the storage. It is usually bigger than L2. This computation that Garg must do and does by standard methods can be done more efficiently by using our present invention.

The Examiner relies upon secondary reference Alpatov, et al., for reasons unrelated to this aspect of the present invention, so that Alpatov fails to overcome this basic deficiency of Garg. It is noted that co-inventor Gunnels of the present invention is one of the seven authors of this secondary reference Alpatov and respectfully disagrees with the Examiner's position.

Hence, turning to the clear language of the claims, in Garg, there is no teaching or suggestion of: ".... streaming data from two of three matrices involved in processing said linear algebra subroutine to a first matrix such that submatrix data of said two matrices residing in a higher level cache or in a memory is streamed to submatrix data of said first matrix residing in

said cache ….", as required by independent claim 1. The remaining independent claims have similar language.

Claim 2 specifically defines how data is passed from L2 into L1 and is clearly patentable over the description in column 7 of Garg, as described above.

Relative to claims 3 and 4, the description in Garg at column 7, lines 38-55, and column 12, lines 17-39, is about prior-art cache blocking and saving storage. In contrast, claim 3 describes choosing an argument that is small relative to the other two, in order to determine which sub-matrix data of a given matrix should be cache resident for the streaming process. Streaming requires that the other two arguments be as large or larger (conservation of matrix data). Applicants submit that the description in Garg has nothing to do with satisfying the plain meaning of the claim language of these claims.

Relative to the rejection based upon secondary reference Aplatov, the present invention is not concerned with Parallel Linear Algebra Libraries. Rather, the present invention addresses serial libraries (e.g., LAPACK). Moreover, primary reference Garg only refers to sparse Cholesky Factorization via a supernodal approach, and part of this approach uses standard state of the art cache blocking techniques. Therefore, Applicants submit that Alpatov is non-analogous to Garg and that both Garg and Alpatov are non-analogous to the present invention.

Moreover, relative to claim 8, this claim is referring to BLAS kernels, whereas Garg is referring to BLAS DGEMM. This comparison of kernels versus DGEMM and also many DLA algorithms versus a single sparse Cholesky Factorization algorithm also confirms that Garg is non-analogous.

Therefore, Applicants submit that there are elements of the claimed invention that are not taught or suggest by Garg. Therefore, the Examiner is respectfully requested to withdraw this rejection.

IV. FORMAL MATTERS AND CONCLUSION

Minor errors have been corrected in the disclosure to update the information for the listing of related co-pending applications.

In view of the foregoing, Applicants submit that claims 1-20, all the claims presently pending in the application, are patentably distinct over the prior art of record and are in condition for allowance. The Examiner is respectfully requested to pass the above application to issue at the earliest possible time.

Should the Examiner find the application to be other than in condition for allowance, the Examiner is requested to contact the undersigned at the local telephone number listed below to discuss any other changes deemed necessary in a telephonic or personal interview.

The Commissioner is hereby authorized to charge any deficiency in fees or to credit any overpayment in fees to Assignee's Deposit Account No. 50-0510.

Respectfully Submitted,

Date: March 19, 2007

Frederick E. Cooperrider
Registration No. 36,769

McGinn Intellectual Property Law Group, PLLC
8321 Old Courthouse Road, Suite 200
Vienna, VA 22182-3817
(703) 761-4100
Customer No. 21254